Cahier des charges
Développement de l'outil Lifemap pour visualiser tout type de jeu de données

Étudiants : Cassandre HÉRITIER—TELLIER, Thomas CATEL, Joe UEDA
Chef De Projet : Damien DE VIENNE
Encadrant : Laurent JACOB

9 octobre 2020

Table des matières

Master de Bioinformatique, première année

1 Présentation du projet

1.1 Contexte

En phylogénétique, la visualisation des espèces qu'elles soient vivantes ou éteintes et de leurs relations sous forme d'un arbre facilite la compréhension des données et des résultats obtenus en génétique, que ce soit par exemple le taux d'expression d'un gène ou le nombre de séquences microsatellite dans un génome. Les corrélations, les hypothèses d'évolution ou toute autre analyse sont bien plus abordables par la visualisation des données à travers l'arbre phylogénétique. Un outil développé par le chef de projet Damien de Vienne, Lifemap, permet une telle visualisation sur l'intégralité de l'arbre du vivant, en ligne et accessible par tous. C'est-à-dire qu'il est possible de se promener à travers l'arbre du vivant, raciné par un ancêtre commun à tous, parmi toutes les taxons qui le compose (espèce, genre, famille, etc.) et de visualiser les liens évolutifs. Il s'agit d'un outil de navigation tel qu'on l'utilise couramment avec OpenStreetMap ou Google Maps par exemple, le principe est le même. Plusieurs versions de Lifemap existent : Lifemap classique, Lifemap-fr et Virusmap. La base de données en ligne exploitable pour apporter des données taxonomiques à Lifemap est celle de SOLR. Une autre version existe, Lifemap NCBI, où l'on peut visualiser des ronds oranges dont la taille est proportionnelle au nombre de génomes séquencés dans NCBI. Cette information a été récupérée en amont et est stockée aussi dans la base SOLR. Voir le lien suivant : `http://lifemap.univ-lyon1.fr/`.

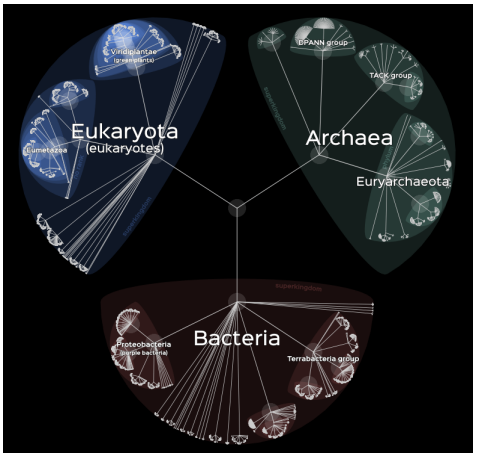

FIGURE 1 – Visualisation de l'arbre du vivant avec Lifemap

1.2 Objectifs

Le but de ce projet est de permettre à un utilisateur quelconque d'être en mesure de visualiser son jeu de données quel qu'il soit, en relation à un lot d'espèces choisi, à travers l'outil Lifemap existant. Ce projet doit aboutir à la création d'un package R, un langage de programmation, qui ira se greffer à l'outil Lifemap. L'objectif premier est de développer une version dédiée à un public ayant une compétence minimum en programmation. Au-delà de cet objectif, il est envisageable de développer une version Shiny de ce package, plus ergonomique, pour que le public sans compétence en programmation puisse aussi profiter de l'outil. Certaines fonctionnalités optionnelles seront développées ou non en fonction de leur complexité et du temps disponible au cours du projet.

1.3 Description de l'existant

Tout d'abord il est nécessaire de connaître l'ensemble des fonctionnalités déjà existantes autour de l'outil Lifemap afin de ne pas développer ce qui a été fait et de gagner du temps. En effet un package *leaflet* existe pour apporter un fond de carte et y ajouter des marqueurs donnés. Ce package écrit en R fait appel au package *leaflet* écrit en JavaScript déjà existant. Il s'agit d'un wrapper qui permet

d'appeler plusieurs librairies Javascript ce qui crée des cartes interactives. Ce package est adapté pour une utilisation sous R (*leaflet* for R : `https://rstudio.github.io/leaflet/`), ce qui facilite la manipulation des fichiers de format *data.frame*. Une documentation très complète accompagne le package, ce qui offre une bonne direction pour les étudiants au cours du projet. De plus un sujet de Travaux Pratiques est enseigné par Damien de Vienne aux étudiants du Master de Bioinformatique de deuxième année, ceci constitue une ressource non négligeable d'informations et d'exercices à maîtriser pour bien démarrer le projet.

1.4 Critères d'acceptabilité du produit

Le projet sera considéré comme fini et réussi si un utilisateur quelconque arrive sans trop de peine à visualiser ses données sur l'arbre du vivant que compose Lifemap. Les critères d'acceptabilité du produit sont les suivants :

- Le rendu est un package R, documenté, qui passe les contrôles automatiques de validation des packages et permet au minimum à un utilisateur de réaliser les actions listées dans les fonctionnalités primaires.
- Les données sont marquées donc visibles clairement et sans ambiguïtés par l'utilisateur sur l'arbre du vivant, et le marqueur cliquable est capable d'afficher les informations principales sur le taxon.
- Les données propres aux espèces sélectionnées correspondent aux bonnes espèces et aucune donnée ou espèce n'est oubliée.
- Le zoom / dé-zoom sur l'arbre ne pose pas de problèmes, les informations des marqueurs se regroupent au noeud parent ou sont placés de façon cohérente et il n'y a pas de bugs ou de ralentissement trop important.
- Tous les formats dans lesquels les données sont apportées par un utilisateur peuvent être traités et utilisés.

2 Expression des besoins

2.1 Fonctionnalités primaires

Ce qui est attendu en priorité dans ce projet concerne les points suivants, exprimant les besoins fonctionnels, qui sont aussi les fonctions à développer en R :

- Mettre les données au format data.frame pour permettre leur utilisation (si les données ne sont pas déjà dans ce format) puis traiter les fichiers de format data.frame afin d'extraire les informations nécessaires pour récupérer les coordonnées des taxons.
- Possibilité de choisir la version de Lifemap à utiliser en fond de carte : Lifemap, Lifemap-fr, Virusmap.
- Possibilité de récupérer dans la base de données en ligne de Lifemap SOLR, pour une liste de taxons, l'ensemble des informations les concernant et nécessaires à leur visualisation sur le fond de carte Lifemap (coordonnées, niveau de zoom, rang taxonomique, nœuds ascendants, etc.).
- Possibilité de lier les données apportées par l'utilisateur et les données récupérées depuis Lifemap.
- Possibilité de visualiser sur le fond de carte Lifemap des données associées aux taxons, et permettre de "résumer" ces données au niveau des noeuds parents des taxons considérés.
- Possibilité de récupérer et de visualiser un sous-arbre à partir d'une liste de taxons (en remontant au MRCA) ou à partir d'un seul taxon (en prenant tous ses descendants).

2.2 Fonctionnalités optionnelles

Les fonctionnalités à développer si les étudiants en ont la possibilité et que le temps le permet sont les suivantes :

— Adapter le package pour afficher une interface utilisable, permettant à un utilisateur sans connaissance en programmation de se servir de l'outil.
— Possibilité de n'afficher lors de l'exploration de l'arbre, que les données réellement visibles à ce niveau de zoom et dans cette zone de l'arbre (utilisation de shiny possible).
— Ajouter des fonctions de mise en évidence, de recherche, et de sélection de données.
— Anticiper les erreurs possibles pour afficher un retour compréhensible au client.
— Créer des fonctions de traduction pour adapter des données mal formatées vers le format supporté par le package.
— Créer une fonction de traduction du nom de l'espèce donnée par l'utilisateur en nom de référence utilisé par le package R (traduire le nom latin en nom commun ou inversement pour homogénéiser et faire comprendre au programme de quel taxon il s'agit).
— Possibilité de re-générer simplement la visualisation (y compris la récupération des données dans la base Lifemap pour que les coordonnées utilisées pour visualiser les données sur ce fond de carte soient en accord avec le fond de carte lui-même).

3 Contraintes

3.1 Délais

Le cahier des charges ci-présent est à rendre le 09 Octobre 2020 à 18h au plus tard. Le projet et sa soutenance doivent aboutir pour le 17 Décembre 2020 selon les modalités de l'université, ce qui laisse aux étudiants deux mois et demi environ pour développer le projet. Toutes les fonctionnalités ne sont pas attendues à ce jour mais les principales doivent fonctionner.

3.2 Autres contraintes

Le projet se fait en partie à distance par les étudiants, qui restent en relation avec le chef de projet et leur encadrant. Pour partager leurs travaux, les étudiants utilisent l'application Discord et développe le projet sur GitHub. Par ailleurs, il faut prendre en compte le fait que la publication d'un package R nécessite de répondre au préalable à plusieurs critères : description, exemple de ligne de code, aide, etc. Le package est ensuite examiné attentivement avant sa publication afin qu'il puisse être utilisé sans crainte de bug.

4 Déroulement du projet

4.1 Planification

Dans un premier temps, en parallèle de l'écriture du cahier des charges, les étudiants s'imprègnent de ce qui existe déjà pour pouvoir se greffer au projet concernant l'outil Lifemap : TP de Master deuxième année à propos de Lifemap, documentation sur le package *leaflet*, réappropriation des outils R et Javascript par exemple. Ensuite la majeure partie du temps sera consacrée au développement du package écrit en R contenant les fonctions qui permettront de réaliser les objectifs. Différents jeux de données seront testés pour voir comment les fonctions se comportent et si leur visualisation se fait sans soucis sur l'arbre du vivant. Enfin, mi-décembre les étudiants présenteront leur projet et ses résultats.

4.2 Plan d'assurance qualité

Pour vérifier la qualité du logiciel à la fin du projet, il est possible de lui faire passer une série de tests. Par exemple, différents jeux de données avec divers formats et lots d'espèces seront apportés au logiciel pour voir comment il se comporte et vérifier les points détaillés précédemment dans les critères d'acceptabilité du produit (1.4). Ce point sera discuté avec le chef de projet une fois les fonctions en place.

4.3 Documentation

Une documentation accompagnera les améliorations apportées au logiciel afin d'utiliser le plus facilement possible le logiciel pour visualiser un jeu de données. Elle sera greffée au package R.

4.4 Diagramme de Gantt

Voici un diagramme qui permet de visualiser en un coup d'œil le déroulement du projet.

TÂCHE	OCTOBRE 2020	NOVEMBRE 2020	DÉCEMBRE 2020
Rédaction du cahier des charges			
Recherche documentaire			
Choix des outils techniques			
Développement du projet			
Rendu du projet			
Présentation du projet			

FIGURE 2 – Diagramme de Gantt